An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano (rgiordan@mit.edu)
Job talk 2021

Dropping data: Motivation

Suppose you're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis (e.g., based the sign / significance of some estimated parameter).

Dropping data: Motivation

Suppose you're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis (e.g., based the sign / significance of some estimated parameter).

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Question: Is the reported interval $-4.55 \pm (5.88)$ a reasonable description of the uncertainty in the estimated efficacy of microcredit?

Can Dropping a Little Data Make a Big Difference?

Do you care whether you can **reverse your conclusion** by removing a **small proportion** of your data?

Not always!

...but sometimes, surely yes.

For example, it often occurs that:

- Policy population is different from analyzed population,
- Small fractions of data are missing not-at-random,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

How do we find influential datapoints?

The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is small.

In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ for $\alpha = 0.0009$.

We provide a fast, automatic approximation based on the **empirical influence function**.

Though we provide finite-sample, non-stochastic accuracy guarantees, there is no need to rely on our theory. A single re-fit provides an exact lower bound on sensitivity.

Can Dropping a Little Data Make a Big Difference?

What causes sensitivity to dropping small fractions of the data?

We consider a number of studies, including seven studies of microcreidt, the Medicaid experiment, a study of cash transfers, and a number of models, including OLS, IV, and a hierachical Bayesian model. Some analyses were robust, and others were not.

What drives the variety of results? We show that sensitivity to dropping small subsets is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N.
- Primarily determined by the **signal to noise** ratio

  ... that is, the ratio of the measured effect size to data variability.

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
"An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?"
https://arxiv.org/abs/2011.14999

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.